DTIC FILE COPY

AD-A214 785

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ~~AFOSR-TR-~~ 89- [redacted] 1655
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): A COMPARISON OF THE FIT OF EMPIRICAL DATA TO TWO LATENT TRAIT MODELS
5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report (Feb. 1, 1978-April 30, 1979)
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Leah R. Hutten
8. CONTRACT OR GRANT NUMBER(s): F49620-78-C-0039
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Laboratory of Psychometric and Evaluative Research, School of Education/University of Massachusetts, Amherst, MA 01003
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Department of the Air Force, Air Force Office of Scientific Research, Bo-ling Air Force Base, DC 20332
12. REPORT DATE: March 1979
13. NUMBER OF PAGES: 27 pages
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.

DTIC ELECTE NOV 28 1989 S B D

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
18. SUPPLEMENTARY NOTES: Paper presented at the annual meeting of AERA, San Francisco, April 1979.
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Few guidelines exist for selecting from the one and three-parameter logistic latent trait models. This study explored fit of empirical data to these two models in terms of degree of violation of model assumptions. Specifically, unidimensionality, guessing, and equality of item discrimination indices were examined. Additionally, fit statistics were explored for data which varied in both sample size and test length. Chi square statistics were used to compare fit of distributions of observed number-right scores to

89 11 27 131

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

AFOSR.TR. 89-1347

20. Abstract (continued)

number right scores predicted from latent trait theory. Using the mean of the conditional distribution of number-right scores for a given ability level as the criterion, the Rasch (one-parameter) model was generally found to be superior in fit to data than the three-parameter model for the five data sets utilized in the study. Fit of data to both models improved as the number of items or persons increased. When short tests were constructed from the data such that item discriminations displayed a broad range, better fit was found for the three-parameter model. Improvement in fit for both models was found for data fulfilling the assumption of unidimensionality. No conclusions were drawn concerning the addition of the guessing parameter in the three-parameter model, since guessing tended to be poorly estimated for the samples of 1000 persons used in this research.

Keywords: Latent trait theory, Cognition, Mathematical models, Psychological tests, Aptitude tests. (SDW)

Accession For		
NTIS GRA&I	☑	
DTIC TAB	☐	
Unannounced	☐	
Justification		
By		
Distribution/		
Availability Codes		
Dist	Avail and/or Special	
A-1		

AFOSR - TR - 79 - [redacted] 1655

A COMPARISON OF THE FIT OF EMPIRICAL DATA TO TWO LATENT TRAIT MODELS[1,2,3]

LEAH R. HUTTEN
UNIVERSITY OF MASSACHUSETTS, AMHERST

---

1. THE PROJECT WAS PERFORMED PURSUANT TO A CONTRACT FROM THE UNITED STATES AIR FORCE OF SCIENTIFIC RESEARCH, HOWEVER, THE OPINIONS EXPRESSED HERE DO NOT NECESSARILY REFLECT THEIR POSITION OR POLICY, AND NO OFFICIAL ENDORSEMENT BY THE AIR FORCE SHOULD BE INFERRED.

2. LABORATORY OF PSYCHOMETRIC AND EVALUATIVE RESEARCH REPORT NO. 92. AMHERST, MA: SCHOOL OF EDUCATION, UNIVERSITY OF MASSACHUSETTS, 1979.

3. A PAPER PRESENTED AT THE ANNUAL MEETING OF AERA, SAN FRANCISCO, APRIL, 1979.

ACKNOWLEDGEMENTS

GRATEFUL ACKNOWLEDGEMENT IS GIVEN TO DR. RONALD K. HAMBLETON, UNIVERSITY OF MASSACHUSETTS, AMHERST, FOR THE DIRECTION GUIDANCE, AND SUPPORT HE HAS GIVEN ME THROUGHOUT THIS PROJECT. SPECIAL THANKS IS ALSO MADE TO DR. ROBERT RENTZ, GEORGIA STATE UNIVERSITY, FOR PROVIDING THE MAJORITY OF DATA USED IN THIS STUDY: STEPHEN IVENS, COLLEGE BOARD, FOR PROVIDING SCHOLASTIC APTITUDE TEST DATA: MARILYN WINGERSKY, EDUCATIONAL TESTING SERVICE, PRINCETON, FOR CONSULTATION AND PROVISION OF THE LOGIST PROGRAM: DR. ERIC GARDNER, SYRACUSE UNIVERSITY, AND THE PSYCHOLOGICAL CORPORATION, NEW YORK CITY, FOR PROVIDING ME WITH STANFORD ACHIEVEMENT TEST DATA: AND TO DORIS PETERSON, UNIVERSITY OF MASSACHUSETTS COMPUTING CENTER, FOR HER ASSISTANCE IN HANDLING MAGNETIC TAPES. MANY HELPFUL COMMENTS WERE RECEIVED FROM DR. H. SWAMINATHAN, UNIVERSITY OF MASSACHUSETTS, AMHERST. IN ADDITION, I AM INDEBTED TO STEPHEN MILES FOR ASSISTANCE IN EDITING AND GENERAL SUPPORT THROUGHOUT THIS STUDY.

A COMPARISON OF THE FIT OF EMPIRICAL DATA TO TWO LATENT TRAIT MODELS

LEAH R. HUTTEN
UNIVERSITY OF MASSACHUSETTS, AMHERST

LATENT TRAIT THEORY HAS SHOWN GREAT PROMISE FOR SOLVING A MULTITUDE OF MEASUREMENT PROBLEMS THAT HAVE NOT BEEN HANDLED ADEQUATELY BY CLASSICAL TEST THEORY. ONE OF THE MOST IMPORTANT GAINS TO BE MADE USING LATENT TRAIT THEORY IS IN THE FIELD OF TEST EQUATING. WITH LATENT TRAIT ABILITY ESTIMATES, IT IS POSSIBLE TO EQUATE TESTS WHICH ARE NOT PARALLEL, AND WHICH DO NOT EVEN CONTAIN THE SAME NUMBER OF ITEMS. THE NATIONAL READING TEST EQUATING STUDY (RENTZ AND BASHAW, 1975) HELPED SPUR INTEREST BY PRACTITIONERS IN LATENT TRAIT ABILITY ESTIMATION. THEORETICALLY IT IS NOW POSSIBLE TO CONDUCT EVALUATIVE STUDIES ON SCHOOL CHILDREN WHO HAVE TAKEN DIFFERENT ACHIEVEMENT TESTS. A SECOND IMPROVEMENT BROUGHT ABOUT THROUGH THE USE OF LATENT TRAIT MODELS OCCURS IN THE FIELD OF TEST DEVELOPMENT. HERE, IT IS POSSIBLE TO DESIGN TESTS AT SPECIFIC DIFFICULTY LEVELS, WHICH CAN BE HIGHLY DISCRIMINATING WITHIN GIVEN ABILITY RANGES. TESTS CAN BE "TAILORED" TO STUDENTS' INDIVIDUAL NEEDS.

BECAUSE MAJOR IMPROVEMENTS IN MEASUREMENT ARE EXPECTED USING LATENT TRAIT THEORY, SCHOOL SYSTEMS AND GOVERNMENT EDUCATIONAL ORGANIZATIONS AROUND THE COUNTRY HAVE SHOWN INCREASED INTEREST IN USING LATENT TRAIT MODELS. THIS INCREASE IN INTEREST IS ALSO ATTRIBUTED TO THE THEORY'S INCREASING ACCEPTANCE BY THE MEASUREMENT COMMUNITY ITSELF, AND FINALLY, TO TECHNOLOGICAL ADVANCES IN BOTH LATENT TRAIT PARAMETER ESTIMATION AND COMPUTER METHODS. ALTHOUGH WE ARE CURRENTLY WITNESSING THE USE OF LATENT TRAIT MODELS IN A

VARIETY OF APPLIED SETTINGS (SEE, FOR EXAMPLE, HAMBLETON ET.AL., 1979; RENTZ AND RENTZ, 1978). MANY BASIC RESEARCH QUESTIONS CONCERNING LATENT TRAIT THEORY HAVE NOT YET BEEN SATISFACTORILY ANSWERED. THE RESEARCH REPORTED IN THIS STUDY WAS DESIGNED TO PROVIDE NEEDED INFORMATION FOR EFFECTIVE APPLICATION OF LATENT TRAIT MODELS BY PRACTITIONERS.

PURPOSE

THE PRIMARY QUESTION ADDRESSED IN THIS STUDY WAS HOW WELL DO EMPIRICAL DATA FIT THE ONE AND THREE-PARAMETER LOGISTIC LATENT TRAIT MODELS,THE MODELS OF MOST CURRENT INTEREST IN THE MEASUREMENT COMMUNITY. ALTHOUGH THERE ARE MANY CLAIMS THAT BOTH ACHIEVEMENT AND APTITUDE DATA FIT RASCH (ONE-PARAMETER) MODELS, AND EQUALLY STRONG CLAIMS CONCERNING FIT OF DATA TO THE THREE-PARAMETER LOGISTIC MODEL, LITTLE RESEARCH HAS ADDRESSED THE QUESTION OF COMPARABLE MODEL FIT. THREE QUESTIONS SEEM ESPECIALLY IMPORTANT:

1.SHOULD THE PRACTITIONER SELECT THE RASCH MODEL WITH ONE TYPE OF DATA, AND THE BIRNBAUM (THREE-PARAMETER) MODEL FOR OTHER KINDS OF DATA?

2.IS THERE IMPROVEMENT IN MODEL-DATA FIT FOUND BY USING THE THREE-PARAMETER MODEL, RATHER THAN THE RASCH MODEL?

3.HOW CAN PRACTITIONERS DETERMINE WHICH TEST MODEL (THE ONE OR THREE-PARAMETER MODEL) BEST SUIT THEIR DATA?

ANSWERS TO THE ABOVE QUESTIONS HAVE BEEN SOUGHT PRIMARILY THROUGH SIMULATION STUDIES. THERE IS INSUFFICIENT EVIDENCE FAVORING ONE OR THE OTHER LATENT TRAIT MODELS FROM RESEARCH USING EMPIRICAL DATA. WHAT FOLLOWS ARE SOME RESULTS

THAT HAVE BEEN ACCUMULATED CONCERNING MODEL FIT. HAMBLETON AND TRAUB (1973) COMPARED THE ONE AND TWO-PARAMETER LOGISTIC MODELS WITH VERBAL AND MATH APTITUDE DATA USING HEURISTIC ESTIMATES OF LATENT TRAIT ITEM PARAMETERS. IMPROVEMENT IN FIT, DEFINED BY A CHI SQUARE TEST BASED ON OBSERVED AND EXPECTED RAW SCORE FREQUENCIES, WAS FOUND FOR THE TWO-PARAMETER MODEL. A RECENT STUDY BY KOCH AND RECKASE (1978) EXPLORED THE FIT OF THE ONE AND THREE-PARAMETER LOGISTIC MODELS FOR APTITUDE AND ACHIEVEMENT TEST DATA USING A MEAN SQUARE DEVIATION STATISTIC. IN THIS STUDY, THE THREE-PARAMETER MODEL CONSISTENTLY FIT DATA BETTER THAN THE ONE-PARAMETER MODEL. UNFORTUNATELY, THE SAMPLING DISTRIBUTION FOR THE MEAN SQUARE DEVIATION STATISTIC IS UNKNOWN, AND THUS THE RESULTS OF THIS STUDY HAVE QUESTIONABLE VALIDITY. RENTZ AND RENTZ (1978) COMPARED THE FIT OF APTITUDE, ACHIEVEMENT, AND CRITERION REFERENCED TEST DATA TO THE RASCH MODEL, USING THE WRIGHT AND PANCHAPAKESAN (1969) FIT STATISTIC. IT WAS REPORTED THAT THE RASCH MODEL ADEQUATELY REPRESENTED THESE THREE DIVERSE KINDS OF DATA. A COMPARISON OF THE ONE, TWO, AND THREE-PARAMETER MODELS WAS CONDUCTED BY HAMBLETON AND COOK (1978) UTILIZING SIMULATED DATA. THIS TECHNIQUE ALLOWED THE RESEARCHERS TO COMPARE ESTIMATED PARAMETERS TO THE TRUE VALUES FROM WHICH THE DATA WERE GENERATED. THESE RESEARCHERS FOUND A SIGNIFICANT IMPROVEMENT BY EMPLOYING THE THREE-PARAMETER LOGISTIC MODEL, ESPECIALLY WITH SHORT TESTS.

THE RESULTS FROM THIS STUDY PROVIDE AN INDICATION OF THE ADEQUACY OF LATENT TRAIT THEORY FOR EXPLAINING TEST BEHAVIOR. THE RESULTS INCLUDE EVIDENCE ON WHICH OF THE ONE OR THREE-PARAMETER LOGISTIC MODELS BEST SUIT VARIOUS TYPES OF DATA. HOPEFULLY, THE INFORMATION PROVIDED HERE CAN SERVE AS A GUIDE FOR PRACTITIONERS IN SELECTING LATENT TRAIT MODELS FOR USE IN TEST CONSTRUCTION AND TEST ANALYSIS.

RESEARCH QUESTIONS

ITEM DISCRIMINATION AND GUESSING

THE RASCH MODEL IS BASED ON THE PREMISES THAT ITEM DISCRIMINATION IS EQUAL FOR ALL ITEMS AND THAT GUESSING DOES NOT OCCUR. TWO QUESTIONS ARISE IN THIS CONNECTION: 1)HOW CAN ONE DETERMINE IF THESE ASSUMPTIONS ARE FULLFILLED IN A DATA SET?, AND 2)CAN DATA FIT THE RASCH MODEL EVEN WHEN THESE ASSUMPTIONS ARE VIOLATED? IT IS DIFFICULT TO ASSUME THAT GUESSING DOES NOT TAKE PLACE ON MULTIPLE CHOICE TESTS, AND YET THE RASCH MODEL IS CONSIDERED ROBUST WITH RESPECT TO THIS CONDITION (MEAD, 1976). A NUMBER OF PRACTICAL PROCEDURES HAVE BEEN SUGGESTED TO DETERMINE THE EXTENT OF GUESSING ON ITEMS. UNFORTUNATELY, MOST METHODS OBSCURE THE POSSIBILITY THAT GUESSING MAY BE AS MUCH PERSON OR ABILITY RELATED AS ITEM RELATED (JENSEMA, 1974). IN THIS CASE, NEITHER THE RASCH OR THE THREE-PARAMETER MODEL WOULD BE AN ADEQUATE DESCRIPTION OF TEST BEHAVIOR. PRACTICAL METHODS ARE UTILIZED IN THIS STUDY TO EXPLORE THE EXTENT OF GUESSING IN A DATA SET.

TWO STRONG POSITIONS ARE TAKEN CONCERNING THE RASCH MODEL ASSUMPTION OF EQUAL ITEM DISCRIMINATION. BIRNBAUM(1968), ROSS (1966), AND HAMBLETON AND TRAUB (1973) FOUND CONSIDERABLE VARIATION IN ITEM DISCRIMINATION FOR EMPIRICAL DATA. NEVERTHELESS, IN STUDIES OF THE RASCH MODEL, RESULTS TYPICALLY SHOW THAT THE MODEL IS FAIRLY ROBUST WITH RESPECT TO VARYING ITEM DISCRIMINATION. FOR EXAMPLE, DINERO AND HAERTEL (1977) EXPLORED SIMULATED DATA IN WHICH CLASSICAL ITEM DISCRIMINATION WAS VARIED UP TO .25 VARIANCE. THEY FOUND NO MAJOR REDUCTION IN FIT TO THE RASCH MODEL. ON THE OTHER HAND, STUDIES BY HAMBLETON AND COOK (1976) AND BY HAMBLETON AND TRAUB (1976), FOUND THE OPPOSITE RESULT, ESPECIALLY WHEN THE RANGE OF VARIATION IN ITEM PARAMETERS WAS LARGE.

THE RANGE OF ITEM DISCRIMINATION CAN BE DETERMINED, TO AN EXTENT, BY EXAMINING CLASSICAL ITEM DISCRIMINATION PARAMETERS. THERE ARE NO REAL GUIDELINES AVAILABLE FOR DETERMINING AT WHAT POINT THE RANGE OF ITEM DISCRIMINATION PARAMETERS IS TOO GREAT TO FIT ASSUMPTIONS OF THE RASCH MODEL. THIS POINT IS ADDRESSED IN THE RESULTS AND CONCLUSIONS OF THIS STUDY.

UNIDIMENSIONALITY

THE ASSUMPTION THAT DATA ARE UNIDIMENSIONAL IS AN ASSUMPTION UNDERLYING NEARLY ALL OF THE POPULAR LATENT TRAIT MODELS. A SINGLE ABILITY, OR LATENT TRAIT, IS ASSUMED TO UNDERLY ITEMS IN A TEST. IN PRACTICE, FEW TEST'S ARE TRULY UNIDIMENSIONAL USING A FACTOR ANALYTIC METHOD. IT IS CUSTOMARY TO FIND LESS THAN 25 % OF A TESTS TOTAL VARIANCE ACCOUNTED FOR BY A FIRST, OR GENERAL, FACTOR . HAMBLETON AND TRAUB (1976) FOUND, WITH ARTIFICIAL DATA, THAT VIOLATION OF THE ASSUMPTION OF UNIDIMENSIONALITY LED TO POOR FIT FOR DATA TO THE RASCH MODEL.

A NUMBER OF TESTS FOR UNIDIMENSIONALITY HAVE BEEN OFFERRED BY VARIOUS RESEARCHERS. LUMSDEN (1961) REVIEWED FIVE METHODS FOR ASSESSING UNIDIMENSIONALITY WITHIN THE CONTEXT OF TEST DEVELOPMENT, AND CONCLUDED THAT FACTOR ANALYSIS IS THE MOST PROMISING METHOD. LATENT TRAIT RESEARCHERS HAVE USED PRINCIPAL COMPONENT ANALYSIS, MAXIMUM LIKELIHOOD FACTOR ANALYSIS, AND PRINCIPAL AXIS COMMON FACTOR ANALYSIS TO DETERMINE UNIDIMENSIONALITY IN THEIR DATA. THERE EXISTS SOME DISAGREEMENT IN THE LITERATURE CONCERNING THE CORRELATION MATRIX THAT IS MOST APPROPRIATE FOR FACTOR ANALYSIS: PHI COEFFICIENTS OR TETRACHORICS. THE LATTER REPRESENTS A MEASURE OF RELATIONSHIP BETWEEN TWO ASSUMED LATENT VARIABLES SCORED DICHOTOMOUSLY. NOT ONLY DOES THIS ASSUMPTION AGREE WITH THE PREMISES OF LATENT TRAIT THEORY, BUT ALSO, USING TETRACHORIC CORRELATIONS IMPROVES THE CHANCES FOR OBTAINING A

FACTOR ANALYTIC SOLUTION. REGARDLESS OF THE STATISTICAL TECHNIQUE USED TO DETERMINE UNIDIMENSIONALITY, ONE PERPLEXING PROBLEM REMAINS: DATA CAN BE UNIDIMENSIONAL FOR ONE SAMPLE AND NOT FOR ANOTHER. CURRENTLY, NO STATISTICAL TECHNIQUE CAN SOLVE THIS PROBLEM. BOTH THE RASCH AND THREE-PARAMETER MODELS ARE INVESTIGATED HERE WITH RESPECT TO HOW WELL THEY FIT DATA OF VARYING DIMENSIONALITY BASED ON A FACTOR ANALYTIC CRITERION.

SAMPLE SIZE AND TEST LENGTH

ONE MAJOR SOURCE OF DISAGREEMENT BETWEEN LATENT TRAIT THEORISTS CONCERNS THE MINIMUM PERSON AND ITEM SAMPLE SIZES NEEDED TO OBTAIN CONSISTENT LATENT TRAIT PARAMETER ESTIMATES. THE LOGIST COMPUTER PROGRAM MANUAL (WOOD, WINGERSKY, AND LORD, 1976) SUGGESTS MINIMUMS OF 40 ITEMS AND 1000 PERSONS. WRIGHT (1977) CONTENDS THAT SMALL SAMPLES (100 PERSONS) ARE SUFFICIENT FOR EFFECTIVE ESTIMATION. THIS STUDY EXPLORES FIT OF SMALL SAMPLE DATA (20 ITEMS, 250 PERSONS) TO THE RASCH AND THREE-PARAMETER MODELS. A CONSIDERABLY MORE EXTENSIVE STUDY OF THIS PROBLEM HAS BEEN PREPARED BY SWAMINATHAN AND GIFFORD (1979).

GOODNESS-OF-FIT

MANY DEFINITIONS FOR GOODNESS-OF-FIT APPEAR IN THE LATENT TRAIT LITERATURE (HAMBLETON, 1979). NOT ONLY DO DEFINITIONS OF FIT VARY FROM AUTHOR TO AUTHOR, BUT METHODS FOR TESTING FIT OF MODELS TO DATA VARY FROM MODEL TO MODEL. MANY OF THE STATISTICAL MEASURES EMPLOYED FOR TESTING GOODNESS-OF-FIT ARE CONSIDERED UNSOUND (BIRNBAUM, 1968). THE CHI SQUARE TEST IS OFTEN UTILIZED FOR GOODNESS-OF-FIT, THOUGH, GIVEN A SUFFICIENT SAMPLE SIZE, MOST DATA WILL BE REJECTED BY THIS MEASURE. NEVERTHELESS, THIS AUTHOR HAS CHOSEN TO EMPLOY CHI SQUARE TEST

STATISTICS IN THIS STUDY. SINCE THE STUDY IS COMPARATIVE IN NATURE, ONLY RELATIVE FIT NEED BE ASSESSED. IN ADDITION, A METHOD WAS NEEDED THAT WOULD BE APPROPRIATE TO BOTH MODELS UNDER STUDY. THE CHI SQUARE STATISTIC MEETS THESE CRITERIA.

METHODOLOGY

DESCRIPTION AND PROCESSING OF TEST DATA

FIVE DATA SETS WERE SELECTED FOR THIS STUDY:

1. CALIFORNIA TEST OF BASIC SKILLS - VOCABULARY SUBTEST, GRADE 10;
2. CALIFORNIIA TEST OF BASIC SKILLS - MATH COMPREHENSION SUBTEST, GRADE 10;
3. SCHOLASTIC APTITUDE TEST - VERBAL, GRADE 12;
4. STANFORD ACHIEVEMENT TEST - VOCABULARY SUBTEST, GRADE 5;
5. STANFORD ACHIEVEMENT TEST - SCIENCE SUBTEST, GRADE 5.

TESTS WERE SELECTED TO COVER A RANGE OF BOTH CONTENT AND GRADE LEVELS. TWO LIMITATIONS WERE PLACED ON DATA SELECTION. FIRST, A MINIMUM SAMPLE SIZE OF 1000 WAS REQUIRED (AT A SINGLE GRADE LEVEL). SECONDLY, THE MINIMUM NUMBER OF ITEMS IN A TEST OR SUBTEST WAS FORTY. EACH OF THE TESTS SELECTED FOR STUDY WAS FOUND TO BE RELATIVELY UNIDIMENSIONAL. IN PILOT ANALYSES, IT WAS FOUND THAT PARAMETER ESTIMATION FOR DATA WHICH IS NOT UNIDIMENSIONAL OFTEN DOES NOT REACH CONVERGENCE WITHIN A REASONABLE TIME LIMIT (400 COMPUTER SECONDS). ANALYSIS OF DATA SETS THAT DO NOT HAVE A DOMINANT SINGLE FACTOR IS PLANNED IN THE NEAR FUTURE.

A FLOW CHART DEPICTING THE DESIGN OF THIS STUDY IS PRESENTED IN FIGURE 1. FOR EACH DATA SET, THE FOLLOWING STEPS

WERE EXECUTED. EACH TEST OR SUBTEST WAS SCORED BY A FORTRAN PROGRAM. THE TETRACHORIC CORRELATION MATRIX WAS OBTAINED AND FACTOR ANALYZED USING A PRINCIPAL COMPONENTS SOLUTION. RESULTS OF THE FACTOR ANALYSIS ARE USED TO CHARACTERIZE DATA IN TERMS OF DIMENSIONALITY. FOLLOWING THE FACTOR ANALYSIS, A RANDOM SAMPLE OF 1000 CASES WAS DRAWN FROM THE TOTAL SAMPLE. THIS SAMPLE WAS RETAINED FOR FURTHER ANALYSIS. CLASSICAL ITEM ANALYSIS WAS PERFORMED TO CHARACTERIZE TESTS IN TERMS OF STANDARD TESTING METHODOLOGY AND TO COMPARE CLASSICAL WITH LATENT TRAIT PARAMETER ESTIMATES. FOR EACH TEST THE AVERAGE, RANGE, AND CONFIDENCE BAND FOR ITEM-TOTAL CORRELATIONS WERE CALCULATED TO EXAMINE THE ASSUMPTION OF EQUAL ITEM DISCRIMINATION. IN ADDITION, CLASSICAL ITEM DIFFICULTIES FOR THE LOWEST DECILE OF EXAMINEES WERE COMPUTED AS AN INDICATOR OF GUESSING ON DIFFICULT ITEMS.

---INSERT FIGURE 1 AROUND HERE-----

IN THE NEXT PHASE OF THE STUDY, ITEM AND ABILITY PARAMETERS WERE ESTIMATED UNDER THE ONE AND THREE-PARAMETER MODELS FOR EIGHT SAMPLING CONDITIONS. TWO SAMPLE SIZES, 250 AND 1000 PERSONS, AND TWO TEST LENGTHS, 20 AND "TOTAL" ITEMS, WERE USED. SAMPLES OF ITEMS WERE SELECTED BY RANDOM METHODS. RANDOM SELECTION OF PERSONS UTILIZED A SPACED SAMPLING TECHNIQUE AFTER VERIFYING THAT THE ORIGINAL SAMPLE OF EXAMINEES WAS NOT ORDERED . PARAMETER ESTIMATION WAS ACCOMPLISHED THROUGH THE LOGIST COMPUTER PROGRAM (WOOD, WINGERSKY, AND LORD, 1976).

SINCE THE INPUT PARAMETER SET FOR EACH LOGIST EXECUTION VARIED GREATLY (OVER 50 PARAMETERS CAN BE SPECIFIED), AN INTERACTIVE TIME-SHARING FORTRAN PROGRAM, LOGPREP, WAS DESIGNED TO CREATE INPUT FILES. FOR MOST THREE-PARAMETER MODEL RUNS THE DEFAULT OPTIONS OF LOGIST WERE USED. THE ONE-PARAMETER MODEL IS ESTIMATED BY FIXING GUESSING AT ZERO AND ITEM DISCRIMINATION AT ONE. OUTPUTS FROM LOGIST ALONG

WITH THE RAW DATA WERE INPUT INTO A FORTRAN PROGRAM, THETITM, TO OBTAIN RAW AND EXPECTED RAW SCORES UTILIZING THE APPROPRIATE ONE OR THREE-PARAMETER ITEM CHARACTERISTIC FUNCTIONS. THE RAW SCORE IS DEFINED AS:

(2.1)

WHERE U =1 IF THE ITEM IS ANSWERED CORRECTLY AND U =0, OTHERWISE. THE EXPECTED RAW SCORE BASED ON LATENT TRAIT THEORY IS:

(2.2)

WHERE P ( ) IS THE PROBABILITY OF A CORRECT RESPONSE ON ITEM G BY PERSONS WITH ABILITY LEVEL THETA, . TO COMPARE OBSERVED AND EXPECTED RAW SCORES (UNDER EACH MODEL) IT WAS NECESSARY TO ROUND EXPECTED RAW SCORES TO THE CLOSEST INTEGER. FINALLY, EXPECTED AND OBSERVED RAW SCORES AND GROUPED RAW SCORE FREQUENCIES WERE OBTAINED USING SPSS. AN INTERACTIVE FORTRAN PROGRAM, CHISQ, WAS USED TO PERFORM CHI SQUARE TESTS FOR EACH MODEL-SAMPLE-TEST LENGTH COMBINATION.THE CHI SQUARE IS DEFINED AS:

(2.3)

O STANDS FOR THE OBSERVED FREQUENCY AND E INDICATES EXPECTED FREQUENCY.

TO ASSESS THE INFLUENCE OF SAMPLE SIZE ON ESTIMATING ITEM PARAMETERS, AN ADDITIONAL LOGIST RUN WAS EXECUTED UNDER THE ASSUMPTIONS OF EACH MODEL. THESE RUNS USED ABILITY ESTIMATES (THETA) FROM THE 1000-PERSON SAMPLE AND RECOMPUTED ITEM PARAMETERS ON A SMALL SAMPLE OF 250 PERSONS. ANALYSIS OF ITEM PARAMETERS WAS THEN ACCOMPLISHED USING THE ADAPT INTERACTIVE STATISTICAL PACKAGE (A TIME-SHARING, APL-BASED STATISTICAL ANALYSIS PACKAGE). ANALYSIS INCLUDED PEARSON AND

SPEARMAN CORRELATIONS BETWEEN SMALL AND LARGE SAMPLE PARAMETERS UNDER THE TWO MODELS, AND IN ADDITION, THE AVERAGE ABSOLUTE DIFFERENCE BETWEEN SMALL AND LARGE SAMPLE PARAMETERS UNDER THE TWO MODELS WAS OBTAINED. A SIMILAR PROCEDURE WAS UTILIZED TO ANALYZE ABILITY ESTIMATES FROM SHORT AND LONG TESTS. IN THIS CASE, ITEM PARAMETERS FOR 20 ITEMS (FROM THE OVERALL "TOTAL" TEST LENGTH ANALYSIS) WERE USED, AND ABILITY ESTIMATES WERE RECOMPUTED FOR THE SHORT TEST UNDER THE ONE AND THREE-PARAMETER MODEL ASSUMPTIONS. THE RESULTING PARAMETER ESTIMATES WERE ANALYZED, AS ABOVE, WITH THE ADAPT STATISTICAL SYSTEM.

FOR EACH LOGIST COMPUTER ESTIMATION COST WAS TALLIED. THE TWO MODELS ARE EXPLORED IN TERMS OF THEIR COMPUTER COSTS. COSTS ARE PRESENTED FOR EACH TEST AND FOR VARIOUS ITEM AND PERSON SAMPLE SIZES.

BECAUSE A NUMBER OF THE RESULTS OF THIS STUDY CONFLICTED WITH THE PREDICTIONS DERIVED FROM THE THEORY OF LATENT TRAITS, ADDITIONAL ANALYSES WERE MADE TO CHECK THE RESULTS. FOUR ADDITIONAL LOGIST ESTIMATIONS WERE EXECUTED ON THE SCHOLASTIC APTITUDE VERBAL SUBTEST. IN EACH CASE A TWENTY (20) ITEM SUBSET OF DATA WAS USED. ONE SUBSET WAS DESIGNED SUCH THAT THE ITEM DISCRIMINATION PARAMETERS WERE EQUAL (A .03 RANGE AROUND THE MEAN POINT-BISERIAL). A SECOND SUBTEST WAS DESIGNED SO THAT THERE RESULTED UNEQUAL ITEM DISCRIMINATIONS (OUTSIDE OF A .1 RANGE ABOUT THE MEAN POINT- BISERIAL). ANALYSES WERE THEN PERFORMED ON THESE DATA TO COMPARE THE ONE AND THREE-PARAMETER MODELS.

RESULTS

FIT OF THE ONE AND THREE-PARAMETER LOGISTIC MODELS

FOR EACH OF THE FIVE DATA SETS, THE EXPECTED RAW SCORE DISTRIBUTION FIT THE OBSERVED RAW SCORE DISTRIBUTION BETTER FOR THE ONE-PARAMETER MODEL THAN FOR THE THREE-PARAMETER MODEL. CHI SQUARE STATISTICS, AVERAGED ACROSS FIVE TESTS, ARE PRESENTED IN TABLE 2. CHI SQUARE STATISTICS FOR EACH INDIVIDAUL TEST ARE PRESENTED IN TABLE 3. THE CHI SQUARE STATISTICS FOR SMALLER SAMPLE SIZES ARE LESS IN MAGNITUDE, AS ONE WOULD EXPECT, ALTHOUGH THERE WERE SOME CONFLICTING RESULTS IN THE DATA. FOR THE ONE-PARAMETER MODEL, THE SHORT TESTS YIELDED BETTER FITS. THE OPPOSITE RESULT HOLDS FOR THE THREE-PARAMETER MODEL. THE DIFFERENCE IN MAGNITUDES FOR THE CHI SQUARES IN TABLE 3 MIGHT BE ATTRIBUTED TO THE WAY IN WHICH THE SCORES WERE GROUPED, ESPECIALLY FOR THE LONG TEST WHICH CONTAINED A VARYING TOTAL NUMBER OF ITEMS. SCORES WERE USUALLY GROUPED INTO SIX CATEGORIES, BUT IN SOME INSTANCES THE LOWEST RAW SCORE GROUP HAD FREQUENCIES TOO LOW FOR COMPUTING THE CHI SQUARE STATISTIC. IN THIS CASE, THE LOWEST TWO SCORE GROUPS WERE COMBINED. ON 20-ITEM TESTS, THE FIRST CATEGORY INCLUDED SCORES 1 THROUGH 4, WHEREAS ALL OTHER CATEGORIES CONTAINED 3 SCORES. ON LONGER TESTS, FIVE OR MORE RAW SCORES COMPOSED EACH GROUPING, WITH THE EXCEPTION OF THE LOWEST AND HIGHEST SCORE GROUPS. THESE CONTAINED FROM SIX TO TWELVE RAW SCORES. ON ANY GIVEN TEST THE GROUPINGS WERE CONSTANT.

---------------INSERT TABLES 2 AND 3 AROUND HERE------

THE VERY HIGH CHI SQUARE STATISTICS CAN ALMOST ALWAYS BE ATTRIBUTED TO LACK OF FIT IN THE LOWEST SCORE GROUPING. THIS EFFECT WAS ESPECIALLY NOTICEABLE FOR THE THREE-PARAMETER MODEL DATA. EVEN WITH THIS SCORE CATEGORY OMITTED, BETTER FIT WAS FOUND FOR THE ONE-PARAMETER MODEL. AN EXCEPTION TO THIS TREND WAS FOUND FOR THE SCIENCE SUBTEST OF THE STANFORD ACHIEVEMENT TEST. HERE, THE FIT TO BOTH MODELS WAS EQUAL. IT SHOULD ALSO BE NOTED THAT THE CRITERION FOR FIT IN THIS STUDY, THE RAW

SCORE, IS A SUFFICIENT STATISTIC FOR THE RASCH MODEL, BUT NOT FOR THE THREE-PARAMETER MODEL. THE RESULTS NEED TO BE CONSIDERED IN VIEW OF THIS FACT.

ITEM DISCRIMINATION, GUESSING, AND UNIDIMENSIONALITY

IT IS IMPOSSIBLE TO OBTAIN A RAW SCORE OF ZERO WITH THE THREE- PARAMETER MODEL IF ANY GUESSING OCCURS. ALTHOUGH LOGIST WAS FAIRLY ACCURATE IN ESTIMATING GUESSING FOR ITEMS FALLING AT THE EXTREMES (NO GUESSING OR MUCH GUESSING), GENERALLY THE GUESSING PARAMETERS WERE UNESTIMABLE. THE ESTIMATION PROCEDURE SETS THE GUESSING PARAMETER TO THE QUANTITY (1/"NCH"-.05) AT THE OUTSET OF ESTIMATION, WHERE NCH IS THE NUMBER OF MULTIPLE CHOICE ALTERNATIVES. IF ESTIMATION OF OTHER PARAMETERS IS STABLE, GUESSING IS ALLOWED TO VARY. THIS WAS NOT USUALLY THE CASE FOR THIS DATA. THE FOLLOWING ARE APPROXIMATE LOWER BOUNDS FOR EXPECTED RAW SCORES UNDER THE THREE-PARAMETER MODEL FOR EACH OF THE FIVE TESTS:

SCHOLASTIC APTITUDE VERBAL = 12.75
CALIFORNIA MATH COMPREHENSION =7.2
CALIFORNIA VOCABULARY = 8.0
STANFORD VOCABULARY = 10.0
STANFORD SCIENCE = 12.0

(THESE LOWER BOUNDS ARE COMPUTED USING THE NUMBER OF ITEMS AND NUMBER OF CHOICES). ALTHOUGH SOME OF THE POOR FIT FOR THE THREE-PARAMETER MODEL CAN BE ATTRIBUTED TO THE LOWEST SCORE GROUP, THE RESULTS WERE STILL RATHER SURPRISING. TWO POSSIBLE EXPLANATIONS EXIST. ONE POSTULATE IS THAT THE DATA CHOSEN FOR STUDY ARE ALL ONE-PARAMETER DATA . A SECOND EXPLANATION IS THAT THERE MAY BE SOME DIFFICULTY IN ESTIMATING PARAMETERS FOR THE THREE-PARAMETER MODEL BECAUSE OF THE ADDITIONAL NUMBER OF UNKNOWN QUANTITIES THAT NEED TO BE ESTIMATED. THE RESULTS ARE MOST LIKELY A COMBINATION OF THESE TWO EXPLANATIONS.

BOTH GUESSING AND ITEM DISCRIMINATION WERE FURTHER INVESTIGATED TO DETERMINE WHETHER THEY HAD BEEN PROPERLY ESTIMATED. TABLE 4 PRESENTS SOME RESULTS CONCERNING THE GUESSING PARAMETER. THE EXTENT OF GUESSING ON EACH TEST WAS DETERMINED BY CALCULATING CLASSICAL ITEM DIFFICULTIES FOR THE 25 % MOST DIFFICULT ITEMS FOR THE LOWEST DECILE OF EXAMINEES BASED ON THE SAMPLE (RAW SCORE CRITERION). ON THIS CRITERION, EACH TEST WAS RATED FOR THE PERCENT OF GUESSING BEHAVIOR DISPLAYED ON HARD ITEMS BY LOW ABILITY EXAMINEES. LATENT TRAIT GUESSING ESTIMATES WERE COMPARED TO THESE VALUES. THE LAST COLUMN OF TABLE 5 INDICATES HOW OFTEN LATENT TRAIT AND CLASSICAL PARAMETERS WERE IN CONCORDANCE, WHICH WAS DEFINED AS THE NUMBER OF TIMES THAT HIGH LATENT TRAIT GUESSING ESTIMATES MATCHED HIGH GUESSING ESTIMATES USING CLASSICAL TEST THEORY INDICATORS. WITH THE EXCEPTION OF THE CALIFORNIA VOCABULARY SUBTEST (WHICH WAS THE SHORTEST AND MOST DIFFICULT TEST), LOGIST WAS QUITE ACCURATE IN PINPOINTING ITEMS AT EITHER EXTREME (MINIMAL OR MAXIMUM GUESSING). GENERALLY THOUGH, THE GUESSING PARAMETER WAS OVERESTIMATED. ALTHOUGH THIS OVERESTIMATION CLEARLY EFFECTED THE LOWEST SCORE GROUP, IN GENERAL, THE EFFECTS OF THIS OVERESTIMATION WERE NOT FOUND ACROSS THE ABILITY DISTRIBUTION. THUS, THE LESS ADEQUATE FIT OF THE DATA TO THE THREE-PARAMETER MODEL CAN NOT BE ATTRIBUTED SOLELY TO OVERESTIMATION OF THE GUESSING PARAMETER.

-------------------INSERT TABLE 4 AROUND HERE------------------

TABLES 5 AND 6 PRESENT RESULTS CONCERNING THE ITEM DISCRIMINATION PARAMETER. THESE RESULTS ARE BASED ON 20-ITEM TESTS CONSTRUCTED TO HAVE VERY DIFFERENT OR VERY SIMILAR ITEM DISCRIMINATIONS (BY CLASSICAL ITEM INDICATORS). IN TABLE 5 CHI SQUARE STATISTICS ARE COMPUTED FOR SIX SCORE GROUPS. IN THIS TABLE WE FIND THAT WHEN THE ITEM DISCRIMINATION

PARAMETERS ARE VERY DIFFERENT, THE THREE-PARAMETER MODEL FITS THE DATA BETTER THAN THE ONE-PARAMETER MODEL. FOR THE CASE OF EQUAL ITEM DISCRIMINATION, THE ONE-PARAMETER MODEL SHOWS BETTER FIT. REGARDLESS OF THE WAY IN WHICH SCORES WERE GROUPED, THE SAME CHI SQUARE TREND WAS FOUND. FROM THESE RESULTS, IT SEEMS PLAUSIBLE TO CONCLUDE THAT ALL OF THE DATA SETS USED IN THIS STUDY HAVE EQUAL ITEM DISCRIMINATIONS. THE AVERAGE CLASSICAL ITEM-TOTAL CORRELATION (POINT-BISERIAL) IS GIVEN FOR EACH DATA SET IN TABLE 6. THE SECOND COLUMN OF THE TABLE SHOWS THE PERCENT OF ITEMS THAT FALL WITHIN THE CONFIDENCE BAND OF THE MEAN POINT-BISERAIL PLUS OR MINUS .1. GENERALLY, THE MAJORITY OF CLASSICAL POINT-BISERAILS ARE QUITE CLOSE IN MAGNITUDE. IT IS SUGGESTED THAT WHEN THE ITEM DISCRIMINATIONS ARE TRULY EQUIVALENT, THE THREE-PARAMETER ESTIMATION PROCEDURE MAY PRODUCE INCONSISTENT ESTIMATES FOR ITEM DISCRIMINATION. RESEARCH CONCURRENT WITH THIS (SWAMINATHAN AND GIFFORD, 1979) HAS INDICATED THAT ITEM DISCRIMINATION TENDS TO BE OVERESTIMATED BY THE MAXIMUM LIKELIHOOD PROCEDURE.

------INSERT TABLES 5 AND 6 AROUND HERE-------------

IT WAS IMPOSSIBLE TO DETERMINE THE INTERRELATIONSHIP BETWEEN GUESSING, ITEM DISCRIMINATION, AND MODEL FIT FOR SPECIFIC DATA SETS IN THIS STUDY. THE TWO SUBTESTS ON WHICH EXAMINEES SHOWED THE MOST GUESSING, ALSO HAD THE NARROWEST RANGE OF ITEM DISCRIMINATIONS. ONE OF THESE, THE STANFORD VOCABULARY, SHOWED CLOSE FIT TO THE RASCH MODEL, AND GOOD FIT TO THE THREE-PARAMETER MODEL AS WELL. THE OTHER, STANFORD SCIENCE, WAS THE SINGLE TEST THAT FIT THE THREE-PARAMETER MODEL AS WELL AS THE RASCH MODEL.

AN EXPLANATION OF MODEL FIT IN TERMS OF UNIDIMENSIONALITY IN THIS STUDY IS CONFOUNDED BY THE FACT THAT TESTS DIFFERED IN BOTH LENGTH AND DIFFICULTY. IT CAN BE SAID, HOWEVER, THAT THE

STANFORD VOCABULARY SUBTEST FIT BOTH MODELS BETTER THAN THE OTHER TESTS, ALTHOUGH THIS TEST WAS NOT THE MOST UNIDIMENSIONAL. TABLE 6 CHARACTERIZES DIMENSIONALITY OF TESTS IN TERMS OF THE FIRST LATENT ROOT FROM THE PRINCIPAL COMPONENT ANALYSIS, AND SHOWS THE VARIANCE ACCOUNTED FOR BY THE FIRST FACTOR. BY THESE CRITERIA, THE TEST WHICH BEST MEETS THE ASSUMPTION OF UNIDIMENSIONALITY IS THE CALIFORNIA MATH TEST. THIS TEST IS ALSO THE EASIEST TEST IN TERMS OF AVERAGE CLASSICAL ITEM DIFFICULTIES. THE RESULTS SHOW THAT THIS TEST FIT BOTH MODELS QUITE WELL. THE CHI SQAURE STATISTIC FOR RASCH MODEL FIT WAS 1.02, THE SECOND BEST FIT FOUND IN THE STUDY.

SAMPLE SIZE

TABLE 7 PROVIDES DATA ON THE ACCURACY OF PARAMETER ESTIMATION FOR SMALL SAMPLES (N=250). THE RESULTS ARE AVERAGED ACROSS THE FIVE TESTS. PEARSON PRODUCT MOMENT CORRELATIONS, SPEARMAN RANK ORDER CORRELATIONS, AND AVERAGE ABSOLUTE DIFFERENCES BETWEEN PARAMETERS ESTIMATED WITH THE 1000 PERSON AND 250 PERSON SAMPLES ARE GIVEN. ALL ESTIMATES WERE FIRST STANDARDIZED TO MEAN ZERO TO OBTAIN THESE RESULTS. ESTIMATES FOR DIFFICULTY ARE QUITE ACCURATE IN THE SMALL SAMPLE FOR BOTH MODELS. THE SMALL SAMPLE ESTIMATE FOR GUESSING, ALTHOUGH CLOSE IN MAGNITUDE TO THE LARGE SAMPLE ESTIMATE, HAD A LOW CORRELATION WITH THE LARGER SAMPLE ESTIMATE. IT IS APPARENT FROM THIS DATA THAT 250 PERSONS MAY NOT BE A SUFFICIENT SAMPLE SIZE UPON WHICH TO ESTIMATE GUESSING. IN FACT, EVEN IN THE 1000-PERSON SAMPLE, THE MAJORITY OF GUESSING PARAMETERS FOR THIS DATA REMAINED UNESTIMATED BY THE MAXIMUM LIKELIHOOD METHOD. ESTIMATION OF ITEM DISCRIMINATION IN THE 250 PERSON SAMPLE IS RELATIVELY CONSISTENT WITH 1000 PERSON ESTIMATE. BUT, BY THE AVERAGE ABSOLUTE DEVIATION CRITERION, THIS SMALL SAMPLE ESTIMATE

FAIRED LESS WELL THAN EITHER GUESSING OR DIFFICULTY. IT APPEARS THAT WHEN DISCRIMINATION IS POORLY ESTIMATED, ALL OTHER ESTIMATES ARE EFFECTED. THEREFORE, THE DIFFICULTY PARAMETERS IN THE THREE-PARAMETER CASE DO NOT APPEAR TO BE ESTIMATED AS EFFECTIVELY WITH SMALL SAMPLES AS IN THE ONE-PARAMETER CASE.

--------INSERT TABLE 7 AROUND HERE-------------------

TEST LENGTH

TEST LENGTH WAS EXAMINED TO DETERMINE WHETHER LATENT TRAIT THEORY CAN BE APPLIED TO SHORT TESTS (20 ITEMS). TABLE 8 PRESENTS THE RESULTS OF THIS ANALYSIS IN TERMS OF PEARSON AND SPEARMAN CORRELATIONS, AND AVERAGE ABSOLUTE DIFFERENCES BETWEEN SHORT AND LONG TESTS, AVERAGED ACROSS FIVE DATA SETS. FOR BOTH MODELS, ESTIMATES OF ABILITY FROM THE SHORT TEST WERE REASONABLY CONSISTENT WITH ESTIMATES DERIVED FROM THE LONGER TESTS. HERE, AS BEFORE, MORE CONSISTENCY WAS FOUND FOR THE ONE-PARAMETER MODEL.

------INSERT TABLE 8 AROUND HERE------

COSTS

IN ADDITION TO FINDING IMPROVEMENT IN FIT FOR THE ONE-PARAMETER MODEL BY STATISTICAL CRITERIA, THE DATA IN TABLE 9 DEMONSTRATE THAT THE COSTS OF ESTIMATING RASCH PARAMETER VALUES ARE CONSIDERABLY LESS THAN THOSE FOR THE THREE-PARAMETER MODEL. THE COSTS SHOWN IN TABLE 9 ARE AVERAGED ACROSS FIVE TESTS. THIS TABLE ALSO SHOWS THE RELATIONSHIP BETWEEN COMPUTER COSTS FOR LATENT TRAIT ESTIMATES AND THE NUMBER OF PERSONS AND ITEMS ESTIMATED. THESE COSTS ARE BASED ON A CHARGE OF $ 400 PER HOUR. THEY DO NOT REFLECT

AUXILIARY COSTS (DISC STORAGE, MAGNETIC TAPES, DATA PREPARATION, ETC.). ALL OF THE FIGURES IN TABLE 9 ARE BASED ON EXECUTIONS OF LOGIST IN WHICH PERSON AND ITEMS ARE ESTIMATED SIMULTANEOUSLY. TABLES 10 AND 11 SHED DIFFERENT LIGHT ON THE COSTS OF THE ONE AND THREE-PARAMETER MODELS. TABLE 10 INDICATES COMPUTER COSTS AVERAGED OVER FIVE 20-ITEM TESTS WHEN ITEM PARAMETERS ARE KNOWN. THERE IS ESSENTIALLY NO DIFFERENCE BETWEEN THE COSTS OF ESTIMATING ABILITY FOR THE ONE AND THREE-PARAMETER MODELS. SINCE THIS IS THE USUAL MANNER IN WHICH LATENT TRAIT THEORY IS APPLIED, THIS EQUIVALENCE OF COSTS SHOULD BE NOTED BY PRACTITIONERS PLANNING TO USE THESE MODELS. TABLE 10 GIVES COMPUTER COSTS FOR LOGIST RUNS AVERAGED ACROSS FIVE TESTS FOR ESTIMATING ITEM PARAMETERS ON SAMPLES OF 250 PERSONS WHEN ABILITY IS KNOWN . THE COSTS GIVEN FOR THIS STUDY CAN ONLY BE GENERALIZED TO THE LOGIST COMPUTER PROGRAM AND DO NOT APPLY TO COMPARISONS WITH OTHER ESTIMATION ROUTINES. IF THE ONE-PARAMETER ESTIMATION HAD BEEN EXECUTED ON THE BICAL COMPUTER PROGRAM (WRIGHT AND MEAD, 1976), THE COMPUTER COSTS FOR THE ONE-PARAMETER MODEL WOULD HAVE BEEN CONSIDERABLY LESS. IN THE BICAL PROCEDURE ONE EQUATION IS NEEDED FOR EACH RAW SCORE CATEGORY, WHEREAS IN THE MAXIMUM LIKELIHOOD METHOD, SEPARATE EQUATIONS ARE NEEDED FOR EACH EXAMINEE.

TABLE 14 HIGHLIGHTS COSTS FOR EACH SUBTEST . THERE IS A RELATIONSHIP BETWEEN THE NUMBER OF ITEMS IN A TEST AND ITS COST, BUT THE HIGHER COSTS FOR SOME SUBTESTS CAN ALSO BE ATTRIBUTED TO A LOWER DEGREE OF UNIDIMENSIONALITY.

--------INSERT TABLES 9,10,11 AND 12 AROUND HERE.------

SUMMARY AND CONCLUSIONS

THE RESULTS OF THIS STUDY INDICATE THAT FOR DATA HAVING

ITEMS EQUAL IN DISCRIMINATION, THE RASCH MODEL PROVIDES BETTER FIT TO EMPIRICAL DATA THAN THE THREE-PARAMETER LOGISTIC MODEL . A PRACTICAL METHOD FOR DETERMINING EQUALITY OF ITEM DISCRIMINATION, USING CLASSICAL POINT-BISERIALS, WAS SUGGESTED. IT WAS ALSO NOTED THAT THE MAXIMUM LIKELIHOOD ESTIMATE OF THE DISCRIMINATION PARAMETER MAY BE INADEQUATE AT THIS TIME. AS IMPROVEMENTS ARE MADE IN THE THREE-PARAMETER ESTIMATION METHODS, A MORE SENSITIVE ESTIMATE OF THIS PARAMETER MAY BE FOUND.

ALTHOUGH THE DATA USED IN THIS STUDY WERE MULTIPLE CHOICE IN NATURE, VIOLATION OF THE "NO GUESSING" ASSUMPTION OF THE RASCH MODEL DID NOT APPEAR TO EFFECT FIT OF THE ONE-PARAMETER MODEL TO DATA. THE MAXIMUM LIKELIHOOD PROCEDURE TENDED TO OVERESTIMATE GUESSING FOR THIS DATA. THIS CAUSED REDUCED MODEL-DATA FIT OF THE THREE-PARAMETER MODEL ESPECIALLY IN THE LOWER ABILITY RANGE. GENERALLY, GUESSING WAS UNESTIMABLE FOR THIS DATA. UNFORTUUNATELY, NO ALTERNATIVE CRITERIA COULD BE FOUND FOR ESTIMATING THE TRUE AMOUNT OF GUESSING. BECAUSE GUESSING AND DISCRIMINATION WERE CONFOUNDED IN THE DATA, IT WAS IMPOSSIBLE TO DETERMINE WHETHER THE GUESSING PARAMETER MIGHT HAVE IMPROVED FIT IN THE THREE-PARAMETER CASE. EMPIRICAL DATA, SUCH AS OPEN-ENDED TEST QUESTIONS, IN WHICH GUESSING IS IMPROBABLE, IS NEEDED TO COMPARE FIT OF THE ONE AND THREE-PARAMETER MODELS. RESEARCH INTO THIS AREA MIGHT BEST BE CONDUCTED THROUGH STUDIES USING SIMULATED DATA. WITH ARTIFICIAL DATA, FACTORS, SUCH AS THOSE CONFOUNDING THE CURRENT RESEARCH, COULD BE CONTROLLED. BETTER ESTIMATES ARE NEEDED FOR BOTH ITEM DISCRIMINATION AND GUESSING IF THE THREE-PARAMETER MODEL IS TO BE USED EFFECTIVELY.

USING A FACTOR ANALYTIC CRITERION, THE DATA USED IN THIS STUDY WERE ALL FOUND TO HAVE ONE GENERAL FACTOR WHICH, IN ALL CASES, ACCOUNTED FOR MORE THAN 20 PERCENT OF THE TEST VARIANCE. THE DATA INDICATE THAT THE MORE A DATA SET MEETS THIS ASSUMPTION, THE LESS TIME IT TAKES TO CONVERGE TO A

SOLUTION BY THE LOGIST PROGRAM. THERE ALSO APPEARED TO BE SOME IMPROVEMENT OF FIT TO BOTH MODELS FOR DATA THAT SHOWED EXTREMELY STRONG FIRST FACTOR VARIANCE. MORE RESEARCH IN THIS AREA IS NEEDED WITH DATA SETS THAT CLEARLY VIOLATE THE ASSUMPTION OF UNIDIMENSIONALITY. IN ADDITION, CRITERIA, OTHER THAN FACTOR ANALYSIS, ARE NEEDED FOR DETERMINING THE EXTENT OF DIMENSIONALITY IN DATA.

ALTHOUGH THE ABILITY ESTIMATES FROM SHORT TESTS WERE REASONABLY GOOD, ITEM ESTIMATES FROM SMALL SAMPLES OF PERSONS TENDED NOT TO BE SO GOOD. THIS RESULT WAS ESPECIALLY APPARENT IN ESTIMATING ITEM DISCRIMINATION FROM SMALL SAMPLES.

WHEN THE LOGIST PROGRAM IS USED WITH KNOWN ITEM PARAMETERS, THE COST OF ESTIMATION IN THE ONE AND THREE-PARAMETER CASES IS EQUIVALENT. IN ESTIMATING ITEM PARAMETERS SIMULTANEOUSLY WITH ABILITY, THE SAVINGS FOUND BY USING THE ONE-PARAMETER MODEL ARE CONSIDERABLE. IT IS DIFFICULT TO COMMENT ON THIS COST DIFFERENTIAL UNTIL IT IS DETERMINED WHETHER THERE ARE OTHER SUBSTANTIAL GAINS TO BE FOUND WITH THE THREE-PARAMETER MODEL.

IN SUMMARY, USING COSTS AND FIT TO TEST SCORE DISTRIBUTIONS AS CRITERIA, THE RASCH MODEL WAS CLEARLY SUPERIOR IN FIT TO EMPIRICAL DATA THAN THE THREE-PARAMETER LOGISTIC MODEL. IT IS IMPORTANT TO POINT OUT THAT OTHER CRITERIA FOR FIT MIGHT HAVE BEEN SELECTED WHICH WOULD HAVE SHOWN BETTER FIT FOR THE THREE-PARAMETER MODEL. FOR EXAMPLE, IF A WEIGHTED RAW SCORE HAD BEEN UTILIZED, RATHER THAN THE SIMPLE RAW SCORE, IMPROVEMENT OF FIT FOR THE THREE-PARAMETER MODEL MIGHT HAVE BEEN SEEN. THE RESULTS ALSO SHOW THAT IN THE CASE WHEN ITEM DISCRIMINATIONS ARE QUITE DISSIMILAR, THE THREE-PARAMETER MODEL DEMONSTRATED SUPERIOR FIT TO THE RASCH MODEL. RESEARCH IS NEEDED TO DETERMINE HOW UNEQUAL ITEM DISCRIMINATION NEED TO BE FOR THE THREE-PARAMETER MODEL TO BECOME MORE EFFECTIVE. HERE AGAIN A SIMULATED-DATA STUDY ,

SIMILAR TO THE ONE PROJECTED ABOVE FOR GUESSING, IS NEEDED IN CONJUNCTION WITH REFINING THE ESTIMATION PROCEDURES.

FINALLY, IT IS IMPORTANT TO POINT OUT THAT THE CONCLUSIONS DRAWN IN THIS PAPER ARE TENTATIVE. THE PROJECT IS IN MIDSTREAM: ONLY HALF OF THE PROJECTED DATA SETS HAVE BEEN ANALYZED TO DATE.

REFERENCES

BIRNBAUM, A. SOME LATENT TRAIT MODELS AND THEIR USE IN INFERRING AN EXAMINEE'S ABILITY. IN F.M. LORD AND M.R. NOVICK, STATISTICAL THEORIES OF MENTAL TESTS. READING, MA: ADDISON-WESLEY, 1968.

DINERO, T.E. AND HAERTEL, E. APPLICABILITY OF THE RASCH MODEL WITH VARYING ITEM DISCRIMINATIONS. APPLIED PSYCHOLOGICAL MEASURMENT, 1977, 1, 581-592.

HAMBLETON, R.K. APPLICATION OF LATENT TRAIT THEORY TO DEVELOPMENT AND USE OF CRITERION REFERENCED TESTS. LABORATORY OF PSYCHOMETRIC AND EVALUATIVE RESEARCH REPORT NO. 91. AMHERST, MA: SCHOOL OF EDUCATION, UNIVERTSITY OF MASSACHUSETTS, 1979.

HAMBLETON, R.K., SWAMINATHAN, H., COOK, L., EIGNOR, D., & GIFFORD, J. DEVELOPMENTS IN LATENT TRAIT THEORY: MODELS, TECHNICAL ISSUES, AND APPLICATIONS. REVIEW OF EDUCATIONAL RESEARCH, 1979 (IN PRESS).

HAMBLETON, R.K. AND COOK, L.L. SOME RESULTS ON THE ROBUSTNESS OF LATENT TRAIT MODELS. PAPER PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, TORONTO, 1978.

HAMBLETON, R.K. AND TRAUB, R.E. ANALYSIS OF EMPIRICAL DATA USING TWO LOGISTIC LATENT TRAIT MODELS. BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY, 1973, 26, 195-211.

HAMBLETON, R.K. AND TRAUB, R.E. SOME EMPITICAL RESULTS ON THE ROBUSTNESS OF THE RASCH TEST THEORY MODEL. LABORATORY OF PSYCHOMETRIC AND EVALUATIVE RESEARCH REPORT NO. 42. AMHERST: THE UNIVERSITY OF MASSACHUSETTS, 1976.

JENSEMA, C.J. AN APPLICATION OF LATENT TRAIT MENTAL TEST THEORY. BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY, 1974, 27, 29-48.

KOCH, B.R. AND RECKASE, M.D. A LIVE TAILORED TESTING COMPARISON OF THE ONE- AND THREE-PARAMETER LOGISTIC MODELS. PAPER PRESENTED AT NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION ANNUAL MEETING, TORONTO, 1978.

LUMSDEN, J. THE CONSTRUCTION OF UNIDIMENSIONAL TESTS. PSYCHOLOGICAL BULLETIN, 1961, 58, 122-131.

MEAD, R. ASSESSING THE FIT OF DATA TO THE RASCH MODEL. PAPER PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, SAN FRANCISCO, 1976.

RENTZ, R.R. AND BASHAW, W.L. EQUATING READING TESTS WITH THE RASCH MODEL: FINAL REPORT. ATHENS, GA: EDUCATIONAL RESEARCH LABORATORY, COLLEGE OF EDUCATION, UNIVERSITY OF GEORGIA, 1975.

RENTZ, R.R. AND RENTZ, C.C. DOES THE RASCH MODEL REALLY WORK? A SYNTHESIS OF THE LITERATURE FOR PRACTITIONERS. MIMEOGRAPHED, 1978.

ROSS, J. AN EMPIRICAL STUDY OF A LOGISTIC MENTAL TEST MODEL. PSYCHOMETRIKA, 1966, 31, 325-340.

SWAMINATHAN, H. AND GIFFORD, J. ESTIMATION OF PARAMETERS IN THE THREE-PARAMETER MODEL. LABORATORY OF PSYCHOMETRIC AND EVALUATIVE RESEARCH REPORT NO. 90. AMHERST, MA: SCHOOL OF EDUCATION, UNIVERSITY OF MASSACHUSETTS, 1979.

WOOD, R.L., WINGERSKY, M.S., AND LORD, F.M. LOGIST: A COMPUTER PROGRAM FOR ESTIMATING ABILITY AND ITEM CHARACTERISTIC CURVE PARAMETERS. RESEARCH MEMORANDUM 76-6. PRINCETON, NJ: EDUCATIONAL TESTING SERVICE, 1976.

WRIGHT, B.D. MISUNDERSTANDING THE RASCH MODEL. JOURNAL OF EDUCATIONAL MEASUREMENT, 1977, 14, 219-225.

WRIGHT, B.D. AND MEAD, R.J. CICAL: CALIBRATING RATING SCALES WITH THE RASCH MODEL. RESEARCH MEMORANDUM NUMBER 23. CHICAGO: STATISTICAL LABORATORY, DEPARTMENT OF EDUCATION, UNIVERSITY OF CHICAGO, 1976.

WRIGHT, B.D., AND PANCHAPAKESAN, N. A PROCEDURE FOR SAMPLE-FREE ITEM ANALYSIS. EDUCATIONAL AND PSYCHOLOGICAL MEASJREMENT, 1969,29, 23-48.

TABLE 1

DESCRIPTION OF TESTS

TEST	ITEMS	MEAN	ST.DEV.	AVG. ITEM DIFF.	COEF. ALPHA	CHOICES	PEOPLE
SAT VERBAL	85	48.114	12.373	.566	.912	5	3000
CTBS MATH	48	30.045	11.634	.625	.944	5	1112
CTBS VOCAB	40	21.572	9.861	.473	.927	4	1112
STEN VOCAB	50	27.814	9.685	.556	.903	4	4160
STEN SCI	60	30.626	12.187	.510	.921	4	4160

TABLE 2

AVERAGE CHI SQUARES ACROSS TESTS
BY MODEL, SAMPLE SIZE, TEST LENGTH

	3-PARAMETER MODEL		1-PARAMETER MODEL	
TEST LENGTH:	LONG	SHORT	LONG	SHORT
SAMPLE SIZE				
1000	7.85	8.51	7.26	5.37
250	4.07	4.94	3.23	2.05
AVG. BY MODEL	6.34		4.48	

TABLE 3

CHI SQUARES BY TEST, MODEL, SAMPLE SIZE, AND TEST LENGTH

		3-PARAMETER MODEL		1-PARAMETER MODEL	
TEST LENGTH:		LONG	SHORT	LONG	SHORT
TEST	SAMPLE SIZE				
SAT VERBAL	1000	17.78	6.36	17.01	7.29
	250	7.81	3.27	4.26	2.63
		8.81		7.80	
CTBS MATH	1000	2.63	3.50	2.81	.44
	250	5.50	1.33	.79	.05
		3.24		1.02	
CTBS VOCAB	1000	5.06	23.74	7.53	7.01
	250	1.37	12.35	7.72	3.70
		10.63		6.49	
STAN VOCAB	1000	3.24	1.18	.08	.10
	250	.46	3.44	.06	.05
		2.08		.07	
STAN SCI	1000	10.53	7.75	8.86	12.03
	250	5.21	4.33	3.34	3.83
		6.96		7.01	

TABLE 4

INVESTIGATION OF GUESSING BY TEST

TEST	ITEMS	25% OF ITEMS	% OF GUESSING ON HARD ITEMS	% OF GOOD ESTIMATES BY LATENT TRAIT ESTIMAT
SAT VERBAL	85	21	43%	76%
CTBS MATH	48	12	50%	83%
CTBS VOCAB	40	10	40%	50%
STAN VOCAB	50	13	70%	77%
STAN SCI	60	14	71%	86%

TABLE 5

CHI SQUARES - 6 SCORE GROUPS
SAT VERBAL 20 ITEM TEST
EQUAL VERSES UNEQUAL ITEM DISCRIMINATION

3-PARAMETER MODEL		1-PARAMETER MODEL	
EQUAL	UNEQUAL	EQUAL	UNEQUAL
43.37	7.85	21.54	9.97

TABLE 6

DESCRIPTION OF TESTS IN TERMS OF MODEL ASSUMPTIONS

TEST	AVG. ITEM-TOTAL CORRELATION	% IN RANGE	FIRST ROOT	% VARIANCE FIRST ROOT	% GUESSING DIFF. ITEMS
SAT VERBAL	.314	78%	19.32	22.7%	43%
CTBS MATH	.491	73%	28.15	58.6%	50%
CTBS VOCAB	.479	70%	20.66	51.7%	40%
STAN VOCAB	.303	82%	14.93	29.9%	70%
STAN SCI	.389	82%	17.23	28.7%	71%

TABLE 7

INFLUENCE OF SAMPLE SIZE ON ITEM PARAMETER ESTIMATION BY MODEL
(250 VERSES 1000 PEOPLE)

	3-PARAMETER MODEL			1-PARAMETER MODEL
PARAMETER:	A	B	C	B
STATISTIC				
PEARSON CORR.	.833	.974	.413	.987
SPEARMAN CORR.	.830	.975	.478	.983
AVG.ABS.DIFF.	.407	.153	.030	.122

TABLE 8

INFLUENCE OF TEST LENGTH ON ABILITY ESTIMATION BY MODEL
(20 VERSES TOTAL ITEMS)

STATISTIC	3-PARAMETER MODEL	1-PARAMETER MODEL
PEARSON CORR.	.866	.923
SPEARMAN CORR.	.918	.926
AVG.ABS.DIFF.	.372	.300

TABLE 9

COMPUTER COSTS BY MODEL

	3-PARAMETER MODEL		1-PARAMETER MODEL	
TEST LENGTH:	LONG	SHORT	LONG	SHORT
SAMPLE SIZE				
1000	$44.40	$19.09	$15.73	$ 5.94
250	$13.45	$ 6.04	$ 5.71	$ 2.55

TABLE 10

COMPUTER COSTS (20 ITEM TEST) WHEN ITEM PARAMETERS ARE KNOWN

3-PARAMETER MODEL	1-PARAMETER MODEL
$3.28	$3.24

TABLE 11

COMPUTER COSTS (250 PEOPLE) WHEN ABILITY SCORES ARE KNOWN

3-PARAMETER MODEL	1-PARAMETER MODEL
$5.46	$3.27

TABLE 12

COMPUTER COSTS BY TEST (N=1000 PERSONS)

	3-PARAMETER MODEL		1-PARAMETER MODEL	
	LONG	SHORT	LONG	SHORT
TEST				
SAT VERBAL	$69.96	*	$26.38	*
CTBS MATH	$31.56	$17.32	$13.66	$ 5.82
CTBS VOCAB	$30.46	$23.18	$10.62	$ 6.23
STAN VOCAB	$30.64	$17.18	$13.00	$ 5.93
STAN SCI	$44.37	$18.39	$15.01	$ 5.79

* NOT RECORDED